CRYSTALLIZATION OF CATHEPSIN S

RELATED APPLICATION

[001] This application claims the benefit of U.S. Provisional Application No. 60/405,423, filed August 23, 2002, which is incorporated herein by reference.

FIELD OF THE INVENTION

[002] The present invention relates to the cysteine protease "cathepsins" of the papain superfamily and more specifically to a particular cathepsin known as Cathepsin S ("CatS"). Provided is CatS in crystalline form, methods of forming crystals comprising CatS, methods of using crystals comprising CatS, a crystal structure of CatS, and methods of using the crystal structure.

BACKGROUND OF THE INVENTION

[003] A general approach to designing inhibitors that are selective for a given protein is to determine how a putative inhibitor interacts with a three dimensional structure of that protein. For this reason it is useful to obtain the protein in crystalline form and perform X-ray diffraction techniques to determine the protein's three dimensional structure coordinates. Various methods for preparing crystalline proteins are known in the art.

[004] Once protein crystals are produced, crystallographic data can be generated using the crystals to provide useful structural information that assists in the design of small molecules that bind to the active site of the protein and inhibit the protein's activity *in vivo*. If the protein is crystallized as a complex with a ligand, one can determine both the shape of the protein's binding pocket when bound to the ligand, as well as the amino acid residues that are capable of close contact with the ligand. By knowing the shape and amino acid residues comprised in the binding pocket, one may design new ligands that will interact favorably with the protein. With such structural information, available computational methods may be used to predict how strong the ligand binding interaction will be. Such methods aid in the design of inhibitors that bind strongly, as well as selectively to the protein.

Express Mailing No. EL978337736US

SUMMARY OF THE INVENTION

[005] The present invention is directed to crystals comprising CatS and particularly crystals comprising CatS that have sufficient size and quality to obtain useful information about the structural properties of CatS and molecules or complexes that may associate with CatS.

[006] In one embodiment, a composition is provided that comprises a protein in crystalline form wherein at least a portion of the protein has 55%, 65%, 75%, 85%, 90%, 95%, 97%, 99% or greater identity with residues SEQ. ID No. 1.

[007] In one variation, the protein has activity characteristic of CatS. For example, the protein may optionally be inhibited by inhibitors of wild type CatS. The protein crystal may also diffract X-rays for a determination of structure coordinates to a resolution of 4Å, 3.5Å, 3.0Å or less.

[008] In one variation, the protein crystal has a crystal lattice in a P4122 space group. The protein crystal may also have a crystal lattice having unit cell dimensions, +/- 5%, of a=b= 85.159Å and c=152.18Å.

[009] The present invention is also directed to crystallizing CatS. The present invention is also directed to the conditions useful for crystallizing CatS. It should be recognized that a wide variety of crystallization methods can be used in combination with the crystallization conditions to form crystals comprising CatS including, but not limited to, vapor diffusion, batch, dialysis, and other methods of contacting the protein solution for the purpose of crystallization.

[0010] The present invention is also directed to crystallizing CatS. The present invention is also directed to the conditions useful for crystallizing CatS. It should be recognized that a wide variety of crystallization methods can be used in combination with the crystallization conditions to form crystals comprising CatS including, but not limited to, vapor diffusion, batch, dialysis, and other methods of contacting the protein solution for the purpose of crystallization.

[0011] In one embodiment, a method is provided for forming crystals of a protein comprising: forming a crystallization volume comprising: a protein wherein at least a portion of the protein has 55%, 65%, 75%, 85%, 90%, 95%, 97%, 99% or greater identity with residues SEQ. ID No. 1; and storing the crystallization volume under conditions suitable for crystal formation.

[0012] In one variation, the crystallization volume comprises the protein in a concentration between 1 mg/ml and 50 mg/ml, and 5-50% w/v of precipitant wherein the

precipitant comprises one or more members of the group comprising PEG having a molecular weight range between 200-20000, 2-methyl-2,4-pentanediol (MPD) and isopropanol, and wherein the crystallization volume has a pH between pH 4 and pH 10.

[0013] The method may optionally further comprise forming a protein crystal that has a crystal lattice in a P4122 space group. The method also optionally further comprises forming a protein crystal that has a crystal lattice having unit cell dimensions, +/- 5%, of a=b= 85.159Å and c=152.18Å. The invention also relates to protein crystals formed by these methods.

[0014] The present invention is also directed to a composition comprising an isolated protein that comprises or consists of one or more of the protein sequence(s) of CatS taught herein for crystallizing CatS. The present invention is also directed to a composition comprising an isolated nucleic acid molecule that comprises or consists of the nucleotides for expressing the protein sequence of CatS taught herein for crystallizing CatS.

[0015] The present invention is also directed to an expression vector that may be used to express the isolated proteins taught herein for crystallizing CatS.

[0016] The present invention is also directed to an expression vector that may be used to express the isolated proteins taught herein for crystallizing CatS. In one variation, the expression vector comprises a promoter that promotes expression of the isolated protein.

[0017] The present invention is also directed to a cell line transformed or transfected by an isolated nucleic acid molecule or expression vector of the present invention.

[0018] The present invention is also directed to structure coordinates for CatS as well as structure coordinates that are comparatively similar to these structure coordinates. It is noted that these comparatively similar structure coordinates may encompass proteins with similar sequences and/or structures, such as other cathepsins. For example, machine-readable data storage media is provided having data storage material encoded with machine-readable data that comprises structure coordinates that are comparatively similar to the structure coordinates of CatS. The present invention is also directed to a machine readable data storage medium having data storage material encoded with machine readable data, which, when read by an appropriate machine, can display a three dimensional representation of all or a portion of a structure of CatS or a model that is comparatively similar to the structure of all or a portion of CatS.

[0019] Various embodiments of machine readable data storage medium are provided that comprise data storage material encoded with machine readable data. The machine readable data

comprises: structure coordinates that have a root mean square deviation equal to or less than the RMSD value specified in Columns 3, 4 or 5 of Table 1 when compared to the structure coordinates of Figure 3, the root mean square deviation being calculated such that the portion of amino acid residues specified in Column 2 of Table 1 of each set of structure coordinates are superimposed and the root mean square deviation is based only on those amino acid residues in the structure coordinates that are also present in the portion of the protein specified in specified in Column 1 of Table 1. The amino acids being overlayed and compared need not to be identical when the RMSD calculation is performed on alpha carbons and main chain atoms but the amino acids being overlayed and compared must have identical side chains when the RMSD calculation is performed on all non-hydrogen atoms.

[0020] For example, in one embodiment where the comparison is based on the 4 Angstrom set of amino acid residues (Column 1) and is based on superimposing alpha-carbon atoms (Column 2), the structure coordinates may have a root mean square deviation equal to or less than 0.65, 0.43, or 0.33 when compared to the structure coordinates of Figure 3.

TABLE 1

AA RESIDUES TO USE TO PERFORM RMSD COMPARISON	PORTION OF EACH AA RESIDUE USED TO PERFORM RMSD COMPARISON	RMSD VALUE LESS THAN OR EQUAL TO		
Table 2 (4 Angstrom set) (relative to E64)	alpha-carbon atoms[1]	0.65	0.43	0.33
	main-chain atoms[1]	0.63	0.42	0.32
	all non-hydrogen[2]	0.67	0.44	0.33
Table 3 (7 Angstrom set) (relative to E64)	alpha-carbon atoms[1]	0.41	0.27	0.20
	main-chain atoms[1]	0.40	0.26	0.20
	all non-hydrogen[2]	0.35	0.24	0.18
Table 4 (10 Angstrom set) (relative to E64)	alpha-carbon atoms[1]	0.36	0.24	0.18
	main-chain atoms[1]	0.37	0.25	0.19
	all non-hydrogen[2]	0.25	0.17	0.13
Table 5 (4 Angstrom set) (relative to Trp186)	alpha-carbon atoms[1]	0.19	0.13	0.10
	main-chain atoms[1]	0.19	0.13	0.10
	all non-hydrogen[2]	0.15	0.10	0.07
Table 6 (7 Angstrom set) (relative to Trp186)	alpha-carbon atoms[1]	0.56	0.38	0.28
	main-chain atoms[1]	0.55	0.37	0.27
	all non-hydrogen[2]	0.62	0.41	0.30
Table 7 (10 Angstrom set) (relative to Trp186)	alpha-carbon atoms[1]	0.30	0.20	0.15
	main-chain atoms[1]	0.30	0.20	0.15
	all non-hydrogen[2]	0.39	0.26	0.19
SEQ. ID No. 1	alpha-carbon atoms[1]	0.29	0.19	0.14
	main-chain atoms[1]	0.29	0.19	0.14
	all non-hydrogen[2]	0.35	0.23	0.17

[1]- the RMSD computed between the atoms of all amino acids that are common to both the target and the reference in the aligned and superposed structure. The amino acids need not to be identical.

[2]- the RMSD computed only between identical amino acids, which are common to both the target and the reference in the aligned and superposed structure.

[0021] The present invention is also directed to a three-dimensional structure of all or a portion of CatS. This three-dimensional structure may be used to identify binding sites, to provide mutants having desirable binding properties, and ultimately, to design, characterize, or identify ligands capable of interacting with CatS. Ligands that interact with CatS may be any type of atom, compound, protein or chemical group that binds to or otherwise associates with the protein. Examples of types of ligands include natural substrates for CatS, inhibitors of CatS, and

heavy atoms. The inhibitors of CatS may optionally be used as drugs to treat therapeutic indications by modifying the in vivo activity of CatS.

[0022] In various embodiments, methods are provided for displaying a three dimensional representation of a structure of a protein comprising:

taking machine readable data comprising structure coordinates that have a root mean square deviation equal to or less than the RMSD value specified in Columns 3, 4 or 5 of Table 1 when compared to the structure coordinates of Figure 3, the root mean square deviation being calculated such that the portion of amino acid residues specified in Column 2 of Table 1 of each set of structure coordinates are superimposed and the root mean square deviation is based only on those amino acid residues in the structure coordinates that are also present in the portion of the protein specified in specified in Column 1 of Table 1;

computing a three dimensional representation of a structure based on the structure coordinates; and

displaying the three dimensional representation.

[0023] The present invention is also directed to a method for solving a three-dimensional crystal structure of a target protein using the structure of CatS.

[0024] In various embodiments, computational methods are provided comprising: taking machine readable data comprising structure coordinates that have a root mean square deviation equal to or less than the RMSD value specified in Columns 3, 4 or 5 of Table 1 when compared to the structure coordinates of Figure 3, the root mean square deviation being calculated such that the portion of amino acid residues specified in Column 2 of Table 1 of each set of structure coordinates are superimposed and the root mean square deviation is based only on those amino acid residues in the structure coordinates that are also present in the portion of the protein specified in specified in Column 1 of Table 1;

computing phases based on the structural coordinates;

computing an electron density map based on the computed phases; and

determining a three-dimensional crystal structure based on the computed electron density map.

[0025] In various embodiments, computational methods are provided comprising: taking an X-ray diffraction pattern of a crystal of the target protein; and computing a three-dimensional electron density map from the X-ray diffraction pattern by molecular replacement, wherein

structure coordinates used as a molecular replacement model comprise structure coordinates that have a root mean square deviation equal to or less than the RMSD value specified in Columns 3, 4 or 5 of Table 1 when compared to the structure coordinates of Figure 3, the root mean square deviation being calculated such that the portion of amino acid residues specified in Column 2 of Table 1 of each set of structure coordinates are superimposed and the root mean square deviation is based only on those amino acid residues in the structure coordinates that are also present in the portion of the protein specified in specified in Column 1 of Table 1.

[0026] These methods may optionally further comprise determining a three-dimensional crystal structure based upon the computed three-dimensional electron density map.

[0027] The present invention is also directed to using a crystal structure of CatS, in particular the structure coordinates of CatS and the surface contour defined by them, in methods for screening, designing, or optimizing molecules or other chemical entities that interact with and preferably inhibit CatS.

[0028] One skilled in the art will appreciate the numerous uses of the inventions described herein, particularly in the areas of drug design, screening and optimization of drug candidates, as well as in determining additional unknown crystal structures. For example, a further aspect of the present invention relates to using a three-dimensional crystal structure of all or a portion of CatS and/or its structure coordinates to evaluate the ability of entities to associate with CatS. The entities may be any entity that may function as a ligand and thus may be any type of atom, compound, protein (such as antibodies) or chemical group that can bind to or otherwise associate with a protein.

[0029] In various embodiments, methods are provided for evaluating a potential of an entity to associate with a protein comprising:

creating a computer model of a protein structure using structure coordinates that comprise structure coordinates that have a root mean square deviation equal to or less than the RMSD value specified in Columns 3, 4 or 5 of Table 1 when compared to the structure coordinates of Figure 3, the root mean square deviation being calculated such that the portion of amino acid residues specified in Column 2 of Table 1 of each set of structure coordinates are superimposed and the root mean square deviation is based only on those amino acid residues in the structure coordinates that are also present in the portion of the protein specified in specified in Column 1 of Table 1;

performing a fitting operation between the entity and the computer model; and

analyzing results of the fitting operation to quantify an association between the entity and the model.

[0030] In other embodiments, methods are provided for identifying entities that can associate with a protein comprising: generating a three-dimensional structure of a protein using structure coordinates that comprise structure coordinates that have a root mean square deviation equal to or less than the RMSD value specified in Columns 3, 4 or 5 of Table 1 when compared to the structure coordinates of Figure 3, the root mean square deviation being calculated such that the portion of amino acid residues specified in Column 2 of Table 1 of each set of structure coordinates are superimposed and the root mean square deviation is based only on those amino acid residues in the structure coordinates that are also present in the portion of the protein specified in specified in Column 1 of Table 1; and

employing the three-dimensional structure to design or select an entity that can associate with the protein; and contacting the entity with a protein wherein at least a portion of the protein has 55%, 65%, 75%, 85%, 90%, 95%, 97%, 99% or greater identity with residues SEQ. ID No. 1.

[0031] In other embodiments, methods are provided for identifying entities that can associate with a protein comprising:

generating a three-dimensional structure of a protein using structure coordinates that comprise structure coordinates that have a root mean square deviation equal to or less than the RMSD value specified in Columns 3, 4 or 5 of Table 1 when compared to the structure coordinates of Figure 3, the root mean square deviation being calculated such that the portion of amino acid residues specified in Column 2 of Table 1 of each set of structure coordinates are superimposed and the root mean square deviation is based only on those amino acid residues in the structure coordinates that are also present in the portion of the protein specified in specified in Column 1 of Table 1; and

employing the three-dimensional structure to design or select an entity that can associate with the protein.

[0032] In other embodiments, methods are provided for identifying entities that can associate with a protein comprising:

computing a computer model for a protein binding pocket, at least a portion of the computer model having a surface contour that has a root mean square deviation equal to or less than a given RMSD value specified in Columns 3, 4 or 5 of Table 1 when the coordinates used to compute the surface contour are compared to the structure coordinates of Figure 3, wherein (a) the root mean square deviation is calculated by the calculation method set forth herein, (b) the portion of amino acid residues associated with the given RMSD value in Table 1 (specified in Column 2 of Table 1) are superimposed according to the RMSD calculation, and (c) the root mean square deviation is calculated based only on those amino acid residues present in both the protein being modeled and the portion of the protein associated with the given RMSD in Table 1 (specified in Column 1 of Table 1); and

employing the computer model to design or select an entity that can associate with the protein; and contacting the entity with a protein wherein at least a portion of the protein has 55%, 65%, 75%, 85%, 90%, 95%, 97%, 99% or greater identity with residues SEQ. ID No. 1.

[0033] In other embodiments, methods are provided for identifying entities that can associate with a protein comprising:

computing a computer model for a protein binding pocket, at least a portion of the computer model having a surface contour that has a root mean square deviation equal to or less than a given RMSD value specified in Columns 3, 4 or 5 of Table 1 when the coordinates used to compute the surface contour are compared to the structure coordinates of Figure 3, wherein (a) the root mean square deviation is calculated by the calculation method set forth herein, (b) the portion of amino acid residues associated with the given RMSD value in Table 1 (specified in Column 2 of Table 1) are superimposed according to the RMSD calculation, and (c) the root mean square deviation is calculated based only on those amino acid residues present in both the protein being modeled and the portion of the protein associated with the given RMSD in Table 1 (specified in Column 1 of Table 1); and

employing the computer model to design or select an entity that can associate with the protein.

[0034] In other embodiments, methods are provided for evaluating the ability of an entity to associate with a protein, the method comprising:

constructing a computer model defined by structure coordinates that have a root mean square deviation equal to or less than the RMSD value specified in Columns 3, 4 or 5 of Table 1

when compared to the structure coordinates of Figure 3, the root mean square deviation being calculated such that the portion of amino acid residues specified in Column 2 of Table 1 of each set of structure coordinates are superimposed and the root mean square deviation is based only on those amino acid residues in the structure coordinates that are also present in the portion of the protein specified in specified in Column 1 of Table 1; and

selecting an entity to be evaluated by a method selected from the group consisting of (i) assembling molecular fragments into the entity, (ii) selecting an entity from a small molecule database, (iii) *de novo* ligand design of the entity, and (iv) modifying a known ligand for CatS, or a portion thereof; performing a fitting program operation between computer models of the entity to be evaluated and the binding pocket in order to provide an energy-minimized configuration of the entity in the binding pocket; and evaluating the results of the fitting operation to quantify the association between the entity and the binding pocket model in order to evaluate the ability of the entity to associate with the binding pocket.

[0035] In other embodiments, methods are provided for evaluating the ability of an entity to associate with a protein, the method comprising:

computing a computer model for a protein binding pocket, at least a portion of the computer model having a surface contour that has a root mean square deviation equal to or less than a given RMSD value specified in Columns 3, 4 or 5 of Table 1 when the coordinates used to compute the surface contour are compared to the structure coordinates of Figure 3, wherein (a) the root mean square deviation is calculated by the calculation method set forth herein, (b) the portion of amino acid residues associated with the given RMSD value in Table 1 (specified in Column 2 of Table 1) are superimposed according to the RMSD calculation, and (c) the root mean square deviation is calculated based only on those amino acid residues present in both the protein being modeled and the portion of the protein associated with the given RMSD in Table 1 (specified in Column 1 of Table 1); and

selecting an entity to be evaluated by a method selected from the group consisting of (i) assembling molecular fragments into the entity, (ii) selecting an entity from a small molecule database, (iii) *de novo* ligand design of the entity, and (iv) modifying a known ligand for CatS, or a portion thereof; performing a fitting program operation between computer models of the entity to be evaluated and the binding pocket in order to provide an energy-minimized configuration of the entity in the binding pocket; and evaluating the results of the fitting operation to quantify the

association between the entity and the binding pocket model in order to evaluate the ability of the entity to associate with the binding pocket.

[0036] In regard to each of these embodiments, the protein may optionally have activity characteristic of CatS. For example, the protein may optionally be inhibited by inhibitors of wild type CatS.

[0037] In another embodiment, a method is provided for identifying an entity that associates with a protein comprising: taking structure coordinates from diffraction data obtained from a crystal of a protein wherein at least a portion of the protein has 55%, 65%, 75%, 85%, 90%, 95%, 97%, 99% or greater identity with residues SEQ. ID No. 1; and performing rational drug design using a three dimensional structure that is based on the obtained structure coordinates.

[0038] The protein crystals may optionally have a crystal lattice with a P4122 space group and unit cell dimensions, +/- 5%, of a=b= 85.159Å and c=152.18Å.

[0039] The method may optionally further comprise selecting one or more entities based on the rational drug design and contacting the selected entities with the protein. The method may also optionally further comprise measuring an activity of the protein when contacted with the one or more entities. The method also may optionally further comprise comparing activity of the protein in a presence of and in the absence of the one or more entities; and selecting entities where activity of the protein changes depending whether a particular entity is present. The method also may optionally further comprise contacting cells expressing the protein with the one or more entities and detecting a change in a phenotype of the cells when a particular entity is present.

BRIEF DESCRIPTION OF THE FIGURES

[0040] Figure 1 illustrates SEQ. ID Nos. 1, 2, 3 and 4 referred to in this application.

[0041] Figure 2 illustrates a crystal of the CatS-E64 complex.

[0042] Figure 3 lists a set of atomic structure coordinates for CatS as derived by X-ray crystallography from a crystal that comprises the protein. The following abbreviations are used in Figure 3: "X, Y, Z" crystallographically define the atomic position of the element measured; "B" is a thermal factor that measures movement of the atom around its atomic center; "Occ" is an occupancy factor that refers to the fraction of the molecules in which each atom occupies the

position specified by the coordinates (a value of "1" indicates that each atom has the same conformation, i.e., the same position, in all molecules of the crystal).

[0043] Figure 4 is a schematic diagram highlighting the secondary structural elements of CatS and the binding region of E64.

[0044] Figure 5A illustrates a surface accessible representation of the molecular surface of CatS showing the long binding pocket with E64 bound in the active site, based on the structure coordinates shown in Figure 3, chain A.

[0045] Figure 5B illustrates key interactions between groups in the binding pockets and the E64 molecule.

[0046] Figure 6 illustrates a system that may be used to carry out instructions for displaying a crystal structure of CatS encoded on a storage medium.

DETAILED DESCRIPTION OF THE INVENTION

[0047] The present invention relates to the cysteine protease "cathepsins" of the papain superfamily and more specifically to a particular cathepsin known as Cathepsin S ("CatS"). More specifically, the present invention relates to CatS in crystalline form, methods of forming crystals comprising CatS, methods of using crystals comprising CatS, a crystal structure of CatS, and methods of using the crystal structure.

[0048] In describing protein structure and function herein, reference is made to amino acids comprising the protein. The amino acids may also be referred to by their conventional abbreviations; A = Ala = Alanine; T = Thr = Threonine; V = Val = Valine; C = Cys = Cysteine; L = Leu = Leucine; Y = Tyr = Tyrosine; I = Ile = Isoleucine; N = Asn = Asparagine; P = Pro = Proline; Q = Gln = Glutamine; F = Phe = Phenylalanine; D = Asp = Aspartic Acid; W = Trp = Tryptophan; E = Glu = Glutamic Acid; M = Met = Methionine; K = Lys = Lysine; G = Gly = Glycine; R = Arg = Arginine; S = Ser = Serine; and H = His = Histidine.

1. CatS

[0049] Cathepsin S (CatS) is a cysteine protease of the papain superfamily. It plays a key role in the generation of a major histocompatibility complex (MHC) class II restricted T-cell response by antigen-presenting cells. Therefore, selective inhibition of this enzyme may be

useful in modulating class II restricted T-cell responses in immune related disorders such as rheumatoid arthritis, multiple sclerosis and extrinsic asthma.

[0050] Cathepsin S is located primarily in lymphatic tissues, the spleen, and in lung macrophages. CatS is a monomeric, 331 amino acids, 37.5 kDa protein. When activated by proteolytic cleavage at low pH, the first 114 residues of the protein are removed, yielding the active, monomeric protein comprising residues 115-331.

[0051] It should be understood that the methods and compositions provided relating to CatS are not intended to be limited to the wild type, full length form of CatS (GenBank Accession Number NP 004070; Wiederanders, B et al "Phylogenetic conservation of cysteine proteases: Cloning and expression of a cDNA coding for human cathepsin S." J. Biol. Chem. 267, 13708-13713). Instead, the present invention also relates to fragments and variants of CatS as described herein.

[0052] In one embodiment, CatS comprises the wild-type form of full length CatS, set forth herein as SEQ. ID No. 1.

[0053] It should be recognized that the invention may be readily extended to various variants of wild-type CatS and variants of fragments thereof. In another embodiment, CatS comprises a sequence that has at least 65% identity, preferably at least 70%, 80%, 90%, 95% or higher identity with SEQ. ID No. 1.

[0054] It is also noted that the above sequences of CatS is also intended to encompass isoforms, mutants and fusion proteins of these sequences. An example of a fusion protein is provided by SEQ. ID No. 3, which includes a 7 residue C-terminal tag (GHHHHHH) that may be used to facilitate purification of the protein. The extended construct is represented in the two sets of structure coordinates shown in Figure 3, represented by chains "A" and "B".

[0055] With the crystal structure provided herein, where amino acid residues are positioned in the structure are now known. As a result, the impact of different substitutions can be more easily predicted and understood.

[0056] For example, based on the crystal structure, applicants have determined that Cat S has an elongated binding pockets capable of binding to an E64 molecule. Figures 5A and 5B illustrate an E64 molecule bound in the CatS binding pocket.

[0057] The amino acids shown in Table 2 were found to be within 4 Angstroms of and therefore close enough to interact with E64. Applicants have also determined that the amino

acids of Table 3 are within 7 Angstroms of E64 and therefore are also close enough to interact with that substrate or analogs thereof. Further it has been determined that the amino acids of Table 4 are within 10 Angstroms of the bound E64.

[0058] The extent of the CatS binding pocket is shown in Table 5 were the listed amino acids are found to be within 4 Angstroms of the CatS binding pocket residue, Trp186. Applicants have also determined that the amino acids of Table 6 are within 7 Angstroms of Trp186 and therefore are also close enough to interact with CatS substrates or analogs thereof. The extent of the CatS binding pocket is further shown in Table 7 were the listed amino acids are found to be within 4 Angstroms of the CatS binding pocket residue, Ser213. Applicants have also determined that the amino acids of Table 8 are within 7 Angstroms of Ser213 and therefore are also close enough to interact with CatS substrates or analogs thereof.

[0059] One or more of these sets of amino acids is preferably conserved in a variant of CatS. Hence, CatS may optionally comprise a sequence that has at least 65% identity, preferably at least 70%, 80%, 90%, 95% or higher identity with SEQ. ID No. 1 where at least the residues shown in Tables 1, 2, 3, 4, 5, 6 and/or 7 are conserved with the exception of 0, 1, 2, 3, or 4 residues. It should be recognized that one might optionally vary some of the binding site residues in order to determine the effect such changes have on structure or activity.

Table 2: CatS binding site residues within 4 Angstroms of E64.

GLN 19	GLY 23	ALA 24
CYS 25	TRP 26	GLY 62
ALA 64	ASN 67	GLY 68
GLY 69	PHE 70	MET 71
GLY 137	VAL 162	ASN 163
HIS 164	GLY 165	

Table 3: CatS binding site residues within 7 Angstroms of E64.

GLN 19	GLY 20	CYS 22
GLY 23	ALA 24	CYS 25
TRP 26	ALA 27	PHE 28
SER 29	TYR 61	GLY 62
ASN 63	ALA 64	GLY 65
CYS 66	ASN 67	GLY 68
GLY 69	PHE 70	MET 71
VAL 136	GLY 137	VAL 138

ALA 140	VAL 162	ASN 163
HUS 164	GLY 165	VAL 166
ASN 184	SER 185	TRP 186
PHE 211		

Table 4: CatS binding site residues within 10 Angstroms of E64.

TYR 18	GLN 19	GLY 20
SER 21	CYS 22	GLY 23
ALA 24	CYS 25	TRP 26
ALA 27	PHE 28	SER 29
ALA 30	VAL 54	SER 57
THR 58	GLU 59	LYS 60
TYR 61	GLY 62	ASN 63
ALA 64	GLY 65	CYS 66
ASN 67	GLY 68	GLY 69
PHE 70	MET 71	THR 72
THR 73	ALA 74	TYR 92
ASP 96	GLU 115	SER 135
VAL 136	GLY 137	VAL 138
ASP 139	ALA 140	ALA 141
PHE 145	PHE 146	GLN 160
ASN 161	VAL 162	ASN 163
HIS 164	GLY 165	VAL 166
LEU 167	ASN 184	SER 185
TRP 186	CYS 206	GLY 207
ILE 208	SER 209	PHE 210
PRO 212	SER 213	

Table 5: CatS binding site residues within 4 Angstroms of Trp186.

LYS 17	TYR 18	GLN 19
PHE 28	ASN 184	SER 185
TRP 186	GLY 187	

Table 6: CatS binding site residues within 7 Angstroms of Trp186.

VAL 16	LYS 17	TYR 18
GLN 19	GLY 20	ALA 24
CYS 25	TRP 26	PHE 28
SER 29	HIS 164	GLY 165
VAL 166	LYS 183	ASN 184
SER 185	TRP 186	GLY 187
HIS 188	PHE 190	GLY 194

Express Mailing No. EL978337736US

Table 7: CatS binding site residues within 4 Angstroms of Ser213.

GLU 115	LEU 116	VAL 134
SER 135	VAL 136	ALA 209
SER 210	PHE 211	PRO 212
SER 213		

Table 8: CatS binding site residues within 7 Angstroms of Ser213.

MET 71	THR 72	PHE 75
THR 114	GLU 115	LEU 116
PRO 117	TYR 118	GLY 119
ARG 120	LEU 124	ALA 127
VAL 134	SER 135	VAL 136
GLY 137	GLY 207	ILE 208
ALA 209	SER 210	PHE 211
PRO 212	SER 213	TYR 214

[0060] With the benefit of the crystal structure and guidance provided by Tables 1, 2, 3, 4, 5, 6 and 7, a wide variety of CatS variants (e.g., insertions, deletions, substitutions, etc.) that fall within the above specified identity ranges may be designed and manufactured utilizing recombinant DNA techniques well known to those skilled in the art, particularly in view of the knowledge of the crystal structure provided herein. These modifications can be used in a number of combinations to produce the variants. The present invention is useful for crystallizing and then solving the structure of the range of variants of CatS.

[0061] Variants of CatS may be insertional variants in which one or more amino acid residues are introduced into a predetermined site in the CatS sequence. For instance, insertional variants can be fusions of heterologous proteins or polypeptides to the amino or carboxyl terminus of the subunits.

[0062] Variants of CatS also may be substitutional variants in which at least one residue has been removed and a different residue inserted in its place. Non-natural amino acids (i.e. amino acids not normally found in native proteins), as well as isosteric analogs (amino acid or otherwise) may optionally be employed in substitutional variants. Examples of suitable substitutions are well known in the art, such as the Glu→Asp, Ser→Cys, Cys→Ser, and His→Ala for example.

[0063] Another class of variants is deletional variants, which are characterized by the removal of one or more amino acid residues from the CatS sequence.

[0064] Other variants may be produced by chemically modifying amino acids of the native protein (e.g., diethylpyrocarbonate treatment that modifies histidine residues). Preferred are chemical modifications that are specific for certain amino acid side chains. Specificity may also be achieved by blocking other side chains with antibodies directed to the side chains to be protected. Chemical modification includes such reactions as oxidation, reduction, amidation, deamidation, or substitution of bulky groups such as polysaccharides or polyethylene glycol.

[0065] Exemplary modifications include the modification of lysinyl and amino terminal residues by reaction with succinic or other carboxylic acid anhydrides. Modification with these agents has the effect of reversing the charge of the lysinyl residues. Other suitable reagents for modifying amino-containing residues include imidoesters such as methyl picolinimidate; pyridoxal phosphate; pyridoxal chloroborohydride; trinitrobenzenesulfonic acid; 0-methylisourea, 2,4-pentanedione; and transaminaseN: talyzed reaction with glyoxylate, and N-hydroxysuccinamide esters of polyethylene glycol or other bulky substitutions.

[0066] Arginyl residues may be modified by reaction with a number of reagents, including phenylglyoxal, 2,3-butanedione, 1,2-cyclohexanedione, and ninhydrin. Modification of arginine residues requires that the reaction be performed in alkaline conditions because of the high pK_a, of the guanidine functional group. Furthermore, these reagents may react with the groups of lysine as well as the arginine epsilon-amino group.

[0067] Tyrosyl residues may also be modified to introduce spectral labels into tyrosyl residues by reaction with aromatic diazonium compounds or tetranitromethane, forming 0-acetyl tyrosyl species and 3-nitro derivatives, respectively. Tyrosyl residues may also be iodinated using ^{125}I or ^{131}I to prepare labeled proteins for use in radioimmunoassays.

[0068] Carboxyl side groups (aspartyl or glutamyl) may be selectively modified by reaction with carbodiimides or they may be converted to asparaginyl and glutaminyl residues by reaction with ammonium ions. Conversely, asparaginyl and glutaminyl residues may be deamidated to the corresponding aspartyl or glutamyl residues, respectively, under mildly acidic conditions. Either form of these residues falls within the scope of this invention.

[0069] Other modifications that may be formed include the hydroxylation of proline and lysine, phosphorylation of hydroxyl groups of seryl or threonyl groups of lysine, arginine and histidine side chains (T. E. Creighton, *Proteins: Structure and Molecular Properties*, W.H.

Freeman & Co., San Francisco, pp. 79-86, 1983), acetylation of the N-terminal amine and amidation of any C-terminal carboxyl group.

[0070] As can be seen, modifications of the nucleic sequence encoding CatS may be accomplished by a variety of well-known techniques, such as site-directed mutagenesis (see, Gillman and Smith, *Gene* 8:81-97 (1979) and Roberts, S. *et al.*, *Nature* 328:731-734 (1987)). When modifications are made, these modifications may optionally be evaluated for there affect on a variety of different properties including, for example, solubility, crystallizability and a modification to the protein's structure and activity.

[0071] In one variation, the variant and/or fragment of wild-type CatS is functional in the sense that the resulting protein is capable of associating with at least one same chemical entity that is also capable of selectively associating with a protein comprising SEQ. ID No. 1 since this common associative ability evidences that at least a portion of the native structure has been conserved. That chemical entity may optionally be E64.

[0072] It is noted the activity of the native protein need not necessarily be conserved. Rather, amino acid substitutions, additions or deletions that interfere with native activity but which do not significantly alter the three-dimensional structure of the domain are specifically contemplated by the invention. Crystals comprising such variants of CatS, and the atomic structure coordinates obtained there from, can be used to identify compounds that bind to the native domain. These compounds may affect the activity or the native domain.

[0073] Amino acid substitutions, deletions and additions that do not significantly interfere with the three-dimensional structure of CatS will depend, in part, on the region where the substitution, addition or deletion occurs in the crystal structure. These modifications to the protein can now be made far more intelligently with the crystal structure information provided herein. In highly variable regions of the molecule, non-conservative substitutions as well as conservative substitutions may be tolerated without significantly disrupting the three-dimensional structure of the molecule. In highly conserved regions, or regions containing significant secondary structure, conservative amino acid substitutions are preferred.

[0074] Conservative amino acid substitutions are well known in the art, and include substitutions made on the basis of similarity in polarity, charge, solubility, hydrophobicity, hydrophilicity and/or the amphipathic nature of the amino acid residues involved. For example, negatively charged amino acids include aspartic acid and glutamic acid; positively charged

amino acids include lysine and arginine; amino acids with uncharged polar head groups having similar hydrophilicity values include the following: leucine, isoleucine, valine; glycine, alanine; asparagine, glutamine; serine, threonine; phenylalanine, tyrosine. Other conservative amino acid substitutions are well known in the art.

[0075] It should be understood that the protein may be produced in whole or in part by chemical synthesis. As a result, the selection of amino acids available for substitution or addition is not limited to the genetically encoded amino acids. Indeed, mutants may optionally contain non-genetically encoded amino acids. Conservative amino acid substitutions for many of the commonly known non-genetically encoded amino acids are well known in the art. Conservative substitutions for other amino acids can be determined based on their physical properties as compared to the properties of the genetically encoded amino acids.

[0076] In some instances, it may be particularly advantageous or convenient to substitute, delete and/or add amino acid residues in order to provide convenient cloning sites in cDNA encoding the polypeptide, to aid in purification of the polypeptide, etc. Such substitutions, deletions and/or additions which do not substantially alter the three dimensional structure of CatS will be apparent to those having skills in the art, particularly in view of the three dimensional structure of CatS provided herein.

2. Cloning, Expression and Purification of CatS

[0077] The gene encoding CatS can be isolated from RNA, cDNA or cDNA libraries. In this case, the portion of the gene encoding amino acid residues 1-331 was isolated and is shown as SEQ. I.D. No. 2.

[0078] Construction of expression vectors and recombinant proteins from the DNA sequence encoding CatS may be performed by various methods well known in the art. For example, these techniques may be performed according to Sambrook et al., Molecular Cloning-A Laboratory Manual, Cold Spring Harbor, N.Y. (1989), and Kriegler, M., Gene Transfer and Expression, A Laboratory Manual, Stockton Press, New York (1990).

[0079] A variety of expression systems and hosts may be used for the expression of CatS. Example 1 provides one such expression system.

[0080] Once expressed, purification steps are employed to produce CatS in a relatively homogeneous state. In general, a higher purity solution of a protein increases the likelihood that

the protein will crystallize. Typical purification methods include the use of centrifugation, partial fractionation, using salt or organic compounds, dialysis, conventional column chromatography, (such as ion exchange, molecular sizing chromatography, etc.), high performance liquid chromatography (HPLC), and gel electrophoresis methods (see, e.g., Deutcher, "Guide to Protein Purification" in Methods in Enzymology (1990), Academic Press, Berkeley, California).

[0081] CatS may optionally be affinity labeled during cloning, preferably with a Glycine-poly-histidine (Gly-His_6) region, in order to facilitate purification. With the use of an affinity label, it is possible to perform a one-step purification process on a purification column that has a unique affinity for the label. The affinity label may be optionally removed after purification. These and other purification methods are known and will be apparent to one of skill in the art.

3. Crystallization & Crystals Comprising CatS

[0082] One aspect of the present invention relates to methods for forming crystals comprising CatS as well as crystals comprising CatS.

[0083] In one embodiment, a method for forming crystals comprising CatS is provided comprising forming a crystallization volume comprising CatS, one or more precipitants, optionally a buffer, optionally an additive and optionally an organic solvent; and storing the crystallization volume under conditions suitable for crystal formation.

[0084] In yet another embodiment, a method for forming crystals comprising CatS is provided comprising forming a crystallization volume comprising CatS in solution comprising the components shown in Table 9; and storing the crystallization volume under conditions suitable for crystal formation.

Table 9

Precipitant 5-50% w/v comprising one or more of any of the PEGs from the 200-20000 molecular weight range, 2-methyl-2,4-pentanediol (MPD) or isopropanol pH pH 4-10. Buffers that may be used include, but are not limited to imidazole, acetate, hepes, citrate, tris, CHES, MES and combinations thereof. Additives 0.1 mM-3 M comprising one or more of benzamidine or 1,2,3-heptanetriol. Protein Concentration 1 mg/ml - 50 mg/ml Temperature 1°C - 25°C

[0085] In yet another embodiment, a method for forming crystals comprising CatS is provided comprising forming a crystallization volume comprising CatS; introducing crystals comprising CatS as nucleation sites, and storing the crystallization volume under conditions suitable for crystal formation.

[0086] Crystallization experiments may optionally be performed in volumes commonly used in the art, for example typically 15, 10, 5, 2 microliters or less. It is noted that the crystallization volume optionally has a volume of less than 1 microliter, optionally 500, 250, 150, 100, 50 or less nanoliters.

[0087] It is also noted that crystallization may be performed by any crystallization method including, but not limited to batch, dialysis and vapor diffusion (e.g., sitting drop and hanging drop) methods. Micro and/or macro seeding of crystals may also be performed to facilitate crystallization.

[0088] It should be understood that forming crystals comprising CatS and crystals comprising CatS according to the invention are not intended to be limited to the wild-type, full

length CatS shown in SEQ. ID No. 1. Rather, it should be recognized that the invention may be extended to various other fragments and variants of wild-type CatS as described above.

[0089] It should also be understood that forming crystals comprising CatS and crystals comprising CatS according to the invention may be such that CatS is complexed with one or more ligands and one or more copies of the same ligand. The ligand used to form the complex may be any ligand capable of binding to CatS. In one variation, the ligand is a natural substrate. In another variation, the ligand is an inhibitor, such as E64.

[0090] In one particular variation, the ligand binds to the first and/or second binding pocket of the protein. Examples of such ligands include, but are not limited to, small molecule inhibitors of CatS such as E64.

[0091] Optionally, the CatS complex may further comprise organics, especially benzamidine which may be introduced in any suitable manner. For example, the organics may be introduced by incubating the desired ligand with a suitable organic such as benzamidine prior to incubation with the CatS protein.

[0092] In one particular embodiment, CatS crystals have a crystal lattice in the $P4_122$ space group. CatS crystals may also optionally have unit cell dimensions, +/- 5%, of a=b= 85.159Å and c=152.18Å.

[0093] CatS crystals also preferably are capable of diffracting X-rays for determination of atomic coordinates to a resolution of greater than 2.2Å.

[0094] Crystals comprising CatS may be formed by a variety of different methods known in the art. For example, crystallizations may be performed by batch, dialysis, and vapor diffusion (sitting drop and hanging drop) methods. A detailed description of basic protein crystallization setups may be found in McRee, D. and David. P., Practical Protein Crystallography, 2nd Ed. (1999), Academic Press Inc. Further descriptions regarding performing crystallization experiments are provided in Stevens, et al. (2000) *Curr. Opin. Struct. Biol.*: 10(5):558-63, and U.S. Patent Nos. 6,296,673, 5,419,278, and 5,096, 676.

[0095] In one variation, crystals comprising CatS are formed by mixing substantially pure CatS with an aqueous buffer containing a precipitant at a concentration just below a concentration necessary to precipitate the protein. One suitable precipitant for crystallizing CatS is polyethylene glycol (PEG), which combines some of the characteristics of the salts and other

organic precipitants (see, for example, Ward et al., *J. Mol. Biol.* 98:161, 1975, and McPherson, *J. Biol. Chem.* 251:6300, 1976.

[0096] During a crystallization experiment, water is removed by diffusion or evaporation to increase the concentration of the precipitant, thus creating precipitating conditions for the protein. In one particular variation, crystals are grown by vapor diffusion in hanging drops or sitting drops. According to these methods, a protein/precipitant solution is formed and then allowed to equilibrate in a closed container with a larger aqueous reservoir having a precipitant concentration for producing crystals. The protein/precipitant solution continues to equilibrate until crystals grow.

[0097] By performing submicroliter volume sized crystallization experiments, as detailed in U.S. Patent No. 6,296,673, effective crystallization conditions for forming crystals of a CatS-E64 complex were obtained. In order to accomplish this, systematic broad screen crystallization trials were performed on a CatS-E64 complex using the sitting drop technique. Over 1000 individual trials were performed in which pH, temperature and precipitants were varied. In each experiment, a 100nL mixture of CatS-E64 complex and precipitant was placed on a platform positioned over a well containing 100□L of the precipitating solution. Precipitate and crystal formation was detected in the sitting drops. Fine screening was then carried out for those crystallization conditions that appeared to produce precipitate and/or crystal in the drops.

[0098] Based on the crystallization experiments that were performed, a thorough understanding of how different crystallization conditions affect CatS crystallization was obtained. Based on this understanding, a series of crystallization conditions were identified that may be used to form crystals comprising CatS. These conditions are summarized in Table 9. A particular example of crystallization conditions that may be used to form crystals diffraction quality crystals of the CatS-E64 complex is detailed in Example 2. Figure 2 illustrates crystals of the CatS-E64 complex formed using the crystallization conditions provided in Table 9.

[0099] One skilled in the art will recognize that the crystallization conditions provided in Table 9 and Example 2 can be varied and still yield protein crystals comprising CatS. For example, it is noted that variations on the crystallization conditions described herein can be readily determined by taking the conditions provided in Table 9 and performing fine screens around those conditions by varying the type and concentration of the components in order to

determine additional suitable conditions for crystallizing CatS, variants of CatS, and ligand complexes thereof.

[00100] Crystals comprising CatS have a wide range of uses. For example, now that crystals comprising CatS have been produced, it is noted that crystallizations may be performed using such crystals as a nucleation site within a concentrated protein solution. According to this variation, a concentrated protein solution is prepared and a crystalline material (microcrystals) is used to 'seed' the protein solution to assist nucleation for crystal growth. If the concentrations of the protein and any precipitants are optimal for crystal growth, the seed crystal will provide a nucleation site around which a larger crystal forms. Given the ability to form crystals comprising CatS according to the present invention, the crystals so formed can be used by this crystallization technique to initiate crystal growth of other CatS comprising crystals, including CatS complexed to other ligands.

[00101] As will be described herein in greater detail, crystals may also be used to perform X-ray or neutron diffraction analysis in order to determine the three-dimensional structure of CatS and, in particular, to assist in the identification of its active site. Knowledge of the binding site region allows rational design and construction of ligands including inhibitors. Crystallization and structural determination of CatS mutants having altered bioactivity allows the evaluation of whether such changes are caused by general structure deformation or by side chain alterations at the substitution site.

4. X-Ray Data Collection and Structure Determination

[00102] Crystals comprising CatS may be obtained as described above in Section 3. As described herein, these crystals may then be used to perform x-ray data collection and for structure determination.

[00103] In one embodiment, described in Example 2, crystals of a CatS-E64 complex were obtained where CatS has the sequence of residues shown in SEQ. ID No. 3. These particular crystals were used to determine the three dimensional structure of CatS. However, it is noted that other crystals comprising CatS including different CatS variants, fragments, and complexes thereof may also be used.

[00104] Diffraction data was collected from cryocooled crystals (100K) of the CatS-E64 complex at the Advanced Light Source beam line 5.0.3 using an ADSC CCD detector. The

diffraction pattern of the CatS-E64 complex displayed symmetry consistent with space group $P4_122$ with unit cell dimensions a=b= 85.159Å and c=152.18Å. Data were collected and integrated to 1.5Å with HKL2000 (Z. Ostwinowski and W. Minor "Processing of X-ray Diffraction Data Collected in Oscillation Mode", Methods in Enzymology, Volume 276: Macromolecular Crystallography, Part A, pages 307-326, 1997, C. W. Carter, Jr. & R. M. Sweet, Eds. Academic Press.).

[00105] All crystallographic calculations were performed using the CCP4 program package (Collaborative Computational Project, N. The CCP4 Suite: Programs for Protein Crystallography. *Acta Cryst.* D50, 760-763 (1994)). The initial phases for the CatS-E64 complex were obtained by the molecular replacement method using the program MOLREP (CCP4). The coordinates of Cathepsin K (PDB code 1AYU) were used as a search model for the solution of the CatS-E64 structure. The highest solution from the translation function was subjected to a rigid body refinement against the maximum likelihood target function as implemented in REFMAC (CCP4). Rigid body refinement was followed by 50 cycles of iterative map/model/phase improvement using ARP_WARP map (Perrakis, A., Morris, R.J. & Lamzin, V.S.) This was followed by alternating cycles of manual rebuilding of the model with Xfit (McRee, D.E. XtalView/Xfit-A versatile program for manipulating atomic coordinates and electron density *J. Struct. Biol.* 125, 156-65 (1999)), ARP_WARP map improvement (Perrakis, A., Morris, R.J. & Lamzin, V.S. Automated protein model building combined with iterative structure refinement) and geometrically restrained refinement against a maximum likelihood target function as implemented in REFMAC (CCP4) until the refinement reached convergence. All stages of model refinement were carried with bulk solvent correction and anisotropic scaling. The data collection and data refinement statistics are given in Table 10.

TABLE 10

Crystal data	
Ligands	E64, water
Space group	$P4_122$
Unit cell dimensions	a=b= 85.159Å and c=152.18Å
Data collection	CatS-E64
X-ray source	ALS 5.0.3
Wavelength [Å]	1.0
Resolution [Å]	50-1.5
Observations (unique)	86434
Redundancy	10
Completeness overall (outer shell)	96% (87%)
I/σ(I) overall (outer shell)	23 (2)
R_{symm} [1] overall (outer shell)	0.066 (0.57)
Refinement	
Reflections used	80501
R-factor	17.1%
R_{free}	19.1%
r.m.s bonds	0.013
r.m.s angles	1.53

[1] $R_{symm} = \Sigma_{hkl}\Sigma_i \left| I(hkl)_i - \langle I(hkl) \rangle \right| / \Sigma_{hkl}\Sigma_i \langle I(hkl)_i \rangle$ over I observations of a reflection hkl

[00106] During structure determination, it was realized that each unit cell comprised two CatS-E64 complexes. Structure coordinates were determined for each of the two complexes in the unit cell. The resultant two sets of structural coordinates from the refinement are presented in Figure 3, as chains A and B. The active site binding pocket is shown in Figures 5A and 5B with a bound E64 molecule. Key interactions between groups in the binding pocket and the E64 molecule are depicted in and described in Figure 5B.

[00107] It is noted that the sequence of the structure coordinates presented in Figures 3 differ in some regards from the sequence shown in SEQ. ID No. 1.

[00108] For some residues, the electron density obtained was insufficient to identify the side chain. As a result, the side chains of these residues were truncated such that a different amino acid is reported. Tables 10 and 11 summarize the differences between SEQ. ID No. 3 and the truncated residues appearing in Figure 3 as chains 'A' and 'B' respectively.

TABLE 11
Truncated Residues in The Structure Coordinates of Figure 3 chain A.

GLU15-ALA	LYS41-ALA	LYS64-ALA
LYS104-ALA	LYS141-ALA	GLU193-ALA

TABLE 12
Truncated Residues in The Structure Coordinates of Figure 3 chain B.

LYS41-ALA	LYS44-ALA	LYS60-ALA
LYS64-ALA	LYS98-ALA	LYS177-ALA

[00109] It is also noted that structure coordinates are not reported for some residues because the electron density obtained was insufficient to identify the position of these residues. For Figure 3, chain A, structure coordinates for residues 220-225 (using numbering from SEQ. No. 3) are not reported. For Figure 3, chain B, structure coordinates for residues 219-225 are not reported.

[00110] Those of skill in the art understand that a set of structure coordinates (such as those in Figure 3) for a protein or a protein-complex or a portion thereof, is a relative set of points that define a shape in three dimensions. Thus, it is possible that an entirely different set of structure coordinates could define a similar or identical shape. Moreover, slight variations in the individual coordinates may have little effect on overall shape. In terms of binding pockets, these variations would not be expected to significantly alter the nature of ligands that could associate with those pockets. The term "binding pocket" as used herein refers to a region of the protein that, as a result of its shape, favorably associates with a ligand

[00111] These variations in coordinates may be generated because of mathematical manipulations of the CatS structure coordinates. For example, the sets of structure coordinates shown in Figure 3 could be manipulated by crystallographic permutations of the structure coordinates, fractionalization of the structure coordinates, application of a rotation matrix,

integer additions or subtractions to sets of the structure coordinates, inversion of the structure coordinates or any combination of the above.

[00112] Alternatively, modifications in the crystal structure due to mutations, additions, substitutions, and/or deletions of amino acids or other changes in any of the components that make up the crystal could also account for variations in structure coordinates. If such variations are within an acceptable standard error as compared to the original coordinates, the resulting three-dimensional shape should be considered to be the same. Thus, for example, a ligand that bound to the active site binding pocket of CatS would also be expected to bind to another binding pocket whose structure coordinates defined a shape that fell within the acceptable error.

[00113] Various computational methods may be used to determine whether a particular protein or a portion thereof (referred to here as the "target protein"), typically the binding pocket, has a high degree of three-dimensional spatial similarity to another protein (referred to here as the "reference protein") against which the target protein is being compared.

[00114] The process of comparing a target protein structure to a reference protein structure may generally be divided into three steps: 1) defining the equivalent residues and/or atoms for the target and reference proteins, 2) performing a fitting operation between the proteins; and 3) analyzing the results. These steps are described in more detail below. All structure comparisons reported herein and the structure comparisons claimed are intended to be based on the particular comparison procedure described below.

[00115] Equivalent residues or atoms can be determined based upon an alignment of primary sequences of the proteins, an alignment of their structural domains or as a combination of both. Sequence alignments generally implement the dynamic programming algorithm of Needleman and Wunsch [*J. Mol. Biol.* 48: 442-453, 1970]. For the purpose of this invention the sequence alignment was performed using the publicly available software program MOE (Chemical Computing Group Inc.) package version 2002.3, as described in the accompanying User's Manual. When using the MOE program, alignment was performed in the sequence editor window using the ALIGN option utilizing the following program parameters: Initial pairwise Build-up: ON, Substitution Matrix: Blosum62, Round Robin: ON, Gap Start: 7, Gap Extend: 1, Iterative Refinement: ON, Build-up: TREE-BASED, Secondary Structure: NONE, Structural Alignment: ENABLED, Gap Start: 1, Gap Extend: 0.1

[00116] Once aligned, a rigid body fitting operation is performed where the structure for the target protein is translated and rotated to obtain an optimum fit relative to the structure of the reference protein. The fitting operation uses an algorithm that computes the optimum translation and rotation to be applied to the moving structure, such that the root mean square deviation of the fit over the specified pairs of equivalent atoms is an absolute minimum. For the purpose of fitting operations made herein, the publicly available software program MOE (Chemical Computing Group Inc.) v. 2002.3 was used.

[00117] The results from this process are typically reported as an RMSD value between two sets of atoms. The term "root mean square deviation" means the square root of the arithmetic mean of the squares of deviations. It is a way to express the deviation or variation from a trend or object. As used herein, an RMSD value refers to a calculated value based on variations in the atomic coordinates of a reference protein from the atomic coordinates of a reference protein or portions of thereof. The structure coordinates for CatS, provided in Figure 3, are used as the reference protein in these calculations.

[00118] The same set of atoms was used for initial fitting of the structures and for computing root mean square deviation values. For example, if a root mean square deviation (RMSD) between Cα atoms of two proteins is needed, the proteins in question should be superposed only on the Cα atoms and not on any other set of atoms. Similarly, if an RMSD calculation for all atoms is required, the superposition of two structures should be performed on all atoms.

[00119] Based on a review of protein structures deposited in the Protein Databank (PDB), 1AYU was identified as having the smallest RMSD values relative to the structure coordinates provided herein. Table 13 below provides a series of RMSD values that were calculated by the above described process using the structure coordinates in Figure 3 as the reference protein and the structure coordinates from PDB code: 1AYU (Human Cathepsin K) as the target protein.

TABLE 13

AA RESIDUES USED TO PERFORM RMSD COMPARISON WITH PDB:1AYU	PORTION OF EACH AA RESIDUE USED TO PERFORM RMSD COMPARISON WITH PDB:1AYU	RMSD [Å]
Table 2 (4 Angstrom set) (relative to E64)	alpha-carbon atoms[1]	1.30
	main-chain atoms[1]	1.27
	all non-hydrogen[2]	1.34
Table 3 (7 Angstrom set) (relative to E64)	alpha-carbon atoms[1]	0.82
	main-chain atoms[1]	0.79
	all non-hydrogen[2]	0.71
Table 4 (10 Angstrom set) (relative to E64)	alpha-carbon atoms[1]	0.73
	main-chain atoms[1]	0.75
	all non-hydrogen[2]	0.51
Table 5 (4 Angstrom set) (relative to Trp186)	alpha-carbon atoms[1]	0.38
	main-chain atoms[1]	0.38
	all non-hydrogen[2]	0.30
Table 6 (7 Angstrom set) (relative to Trp186)	alpha-carbon atoms[1]	1.13
	main-chain atoms[1]	1.09
	all non-hydrogen[2]	1.23
Table 7 (10 Angstrom set) (relative to Trp186)	alpha-carbon atoms[1]	0.61
	main-chain atoms[1]	0.60
	all non-hydrogen[2]	0.78
SEQ. ID No. 1	alpha-carbon atoms[1]	0.58
	main-chain atoms[1]	0.58
	all non-hydrogen[2]	0.70

[1]- the RMSD computed between the atoms of all amino acids that are common to both the target and the reference in the aligned and superposed structure. The amino acids need not to be identical.

[2]- the RMSD computed only between identical amino acids, which are common to both the target and the reference in the aligned and superposed structure.

[00120] It is noted that mutants and variants of CatS as well as other cathepsins are likely to have similar structures despite having different sequences. For example, the binding pockets of these related proteins are likely to have similar contours. Accordingly, it should be recognized that the structure coordinates and binding pocket models provided herein have utility for these other related proteins.

[00121] Accordingly, in one embodiment, the invention relates to data, computer readable media comprising data, and uses of the data where the data comprises all or a portion of the

structure coordinates shown in Figure 3 or structure coordinates having a root mean square deviation (RMSD) equal to or less than the RMSD value specified in Columns 3, 4 or 5 of Table 1 when compared to the structure coordinates of Figure 3, the root mean square deviation being calculated such that the portion of amino acid residues specified in Column 2 of Table 1 of each set of structure coordinates are superimposed and the root mean square deviation is based only on those amino acid residues in the structure coordinates that are also present in the portion of the protein specified in specified in Column 1 of Table 1.

[00122] As noted, there are many different ways to express the surface contours of the CatS structure other than by using the structure coordinates provided in Figure 3. Accordingly, it is noted that the present invention is also directed to any data, computer readable media comprising data, and uses of the data where the data defines a computer model for a protein binding pocket, at least a portion of the computer model having a surface contour that has a root mean square deviation equal to or less than a given RMSD value specified in Columns 3, 4 or 5 of Table 1 when the coordinates used to compute the surface contour are compared to the structure coordinates of Figure 3, wherein (a) the root mean square deviation is calculated by the calculation method set forth herein, (b) the portion of amino acid residues associated with the given RMSD value in Table 1 (specified in Column 2 of Table 1) are superimposed according to the RMSD calculation, and (c) the root mean square deviation is calculated based only on those amino acid residues present in both the protein being modeled and the portion of the protein associated with the given RMSD in Table 1 (specified in Column 1 of Table 1).

5. CatS-E64 Structure

[00123] The present invention is also directed to a three-dimensional crystal structure of CatS. This crystal structure may be used to identify binding sites, to provide mutants having desirable binding properties, and ultimately, to design, characterize, or identify ligands that interact with CatS.

[00124] The three-dimensional crystal structure of CatS may be generated, as is known in the art, from the structure coordinates shown in Figure 3 and similar such coordinates.

[00125] The refined crystal structure of CatS-E64, determined according to the present invention, contains amino acids residues 115-331 as numbered according to SEQ. ID No. 1

(based on the coordinates of Figure 3) and a bound E64 molecule. A total of 627 water molecules were included.

[00126] Figure 4 illustrates a schematic diagram overview of the structure of CatS, highlighting the secondary structural elements of the protein. CatS forms a monomeric structure consisting of two domains. Domain 1 contains three helices and a hydrophobic core, while domain 2 consists of an anti-parallel β-barrel enclosing a hydrophobic core flanked by α-helices on either side of the barrel surface. Three disulfides, two in domain 1 and one in domain 2, play a role in stabilizing the overall structure of the protein. The structure of CatS is very similar to the structures of the plant protein papain (PDB code 9PAP), cathepsin K (PDB code 1MEM), cathepsin H (PDB code 8PCH), and cathepsin L (PDB code 1CS8). Superposition of 200 aligned residues in these structures with the CatS structure results in a root mean square deviation of less than 0.8 Angstroms in the positions of the main chain atoms in all cases.

[00127] Figure 5B shows the detailed interactions between E64 bound in the active site of CatS based on the structure coordinates shown in Figure 3, chain A. The binding pocket of CatS is long and deep, as shown in Figure 5A, and is lined mainly by polar residues. A covalent link is formed between a carbon atom on E64 and the sulfur atom of C25.

6. CatS Binding Pocket and Ligand Interaction

[00128] The term "binding site" or "binding pocket", as the terms are used herein, refers to a region of a protein that, as a result of its shape and surface properties (charge, hydrophobicity, etc.), favorably associates with a ligand or substrate. The term "CatS-like binding pocket" refers to a portion of a molecule or molecular complex whose shape is sufficiently similar to the CatS binding pockets as to bind common ligands. This commonality of shape may be quantitatively defined based on a comparison to a reference point, that reference point being the structure coordinates provided herein. For example, the commonality of shape may be quantitatively defined based on a root mean square deviation (rmsd) from the structure coordinates of the backbone atoms of the amino acids that make up the binding pockets in CatS (as set forth in Figure 3).

[00129] The "active site binding pocket" (Figure 5A) or "active site" of CatS refers to the area on the surface of CatS where the substrate (and an E64 inhibitor molecule) binds. Figures

5A and 5B illustrate E64 bound in the active site of CatS based on the crystal structure of the present invention.

[00130] To date, the active site binding pockets of cathepsins have been actively pursued as targets for the design of small molecule inhibitors. The crystal structure described in the present invention represents the highest resolution, and thus most accurate structure of a member of the cathepsin class (compared to available structures in the public domain), which could be used for the design of more potent inhibitors. A number of key substrate binding and catalytic residues observed in the active site binding pocket of CatS are conserved among all cathepsins. However, sequence variability exists between CatS and other members of the cathepsin protein family, so that the subtle differences in shape and chemical content may be explored to confer specificity of inhibition.

[00131] In resolving the crystal structure of CatS in complex with E64, applicants determined that CatS amino acids in Table 2 (above) are within 4 Angstroms of and therefore close enough to interact with the E64 molecule. Applicants have determined that the amino acids of Table 3 (above) are within 7 Angstroms of bound E64 and therefore are also close enough to interact with that inhibitor or analogs thereof. Applicants have also determined that the amino acids of Table 4 (above) are within 10 Angstroms of bound E64 and therefore are also close enough to interact with that inhibitor or analogs thereof. Applicants have also determined that the amino acids of Table 5 (above) are within 4 Angstroms of Trp186 and therefore are also close enough to interact with substrates/potential inhibitors. Applicants have also determined that the amino acids of Table 6 (above) are within 7 Angstroms of Trp186 and therefore are also close enough to interact with substrates/potential inhibitors. Applicants have also determined that the amino acids of Table 7 (above) are within 4 Angstroms of Ser213 and therefore are also close enough to interact with substrates/potential inhibitors. Applicants have also determined that the amino acids of Table 8 (above) are within 7 Angstroms of Ser213 and therefore are also close enough to interact with substrates/potential inhibitors. The sets of amino acids described in Tables 3 through 7 are preferably conserved in variants of CatS. While it is desirable to largely conserve these residues, it should be recognized however that variants may also involve varying 1, 2, 3, 4 or more of the residues set forth in Tables 1, 2, 3, 4, 5, 6 and 7 in order to evaluate the roles these amino acids play in the binding pocket.

[00132] With the knowledge of the CatS crystal structure provided herein, Applicants define the CatS binding pocket where the relative positioning of the 4, 7, and/or 10 Angstroms sets of amino acids are substantially conserved. Again, it is noted that it may be desirable to form variants where 1, 2, 3, 4 or more of the residues set forth in Tables 1, 2, 3, 4, 5, 6 and 7 are varied in order to evaluate the roles these amino acids play in the binding pockets. Accordingly, any set of structure coordinates for a protein from any source having a root mean square deviation of main-chain atoms of less than 0.3 Å when superimposed on the main-chain atom positions of the corresponding atomic coordinates of either Figure 3, chain A or chain B, for the 4, 7, and/or 10 Angstroms sets of amino acids shall be considered identical. As noted previously, the root mean square deviation is intended to be limited to only those main-chain atoms of amino acid residues that are common to both the protein fragments represented in Figure 3 chain, A or B, and the protein whose structure coordinates are being compared to the coordinates shown in Figure 3 chain, A or B, since the sequence of the protein may be varied somewhat.

[00133] In one embodiment, the invention relates to data, computer readable media comprising data, and uses of the data where the data comprises the structure coordinates shown in Figure 3, chain A or B, or structure coordinates having a root mean square deviation of non-hydrogen atoms of less than 0.8 Å when superimposed on the non-hydrogen atom positions of the corresponding atomic coordinates of Figure 3, chains A or chain B, for the 4, 7, and/or 10 Angstroms sets of amino acids detailed in tables 1, 2, 3, 4, 5, 6 and/or 7. Optionally, the root mean square deviation of non-hydrogen atoms is less than 0.8 Å, 0.7 Å, 0.6 Å, 0.5 Å, 0.4 Å, 0.3 Å, or less.

[00134] It will be readily apparent to those of skill in the art that the numbering of amino acids in other isoforms of CatS may be different than that set forth for CatS. Corresponding amino acids in other isoforms of CatS are easily identified by visual inspection of the amino acid sequences or by using commercially available homology software programs, as further described below.

7. System For Displaying the Three Dimensional Structure of CatS

[00135] The present invention is also directed to machine-readable data storage media having data storage material encoded with machine-readable data that comprises structure coordinates for CatS. The present invention is also directed to a machine readable data storage

media having data storage material encoded with machine readable data, which, when read by an appropriate machine, can display a three dimensional representation of a structure of CatS.

[00136] All or a portion of the CatS coordinate data shown in Figure 3, when used in conjunction with a computer programmed with software to translate those coordinates into the three-dimensional structure of CatS may be used for a variety of purposes, especially for purposes relating to drug discovery. Software for generating three-dimensional graphical representations are known and commercially available. The ready use of the coordinate data requires that it be stored in a computer-readable format. Thus, in accordance with the present invention, data capable of being displayed as the three-dimensional structure of CatS and/or portions thereof and/or their structurally similar variants may be stored in a machine-readable storage medium, which is capable of displaying a graphical three-dimensional representation of the structure.

[00137] For example, in various embodiments, a computer is provided for producing a three-dimensional representation of at least an CatS-like binding pocket, the computer comprising:

machine readable data storage medium comprising a data storage material encoded with machine-readable data, the machine readable data comprising structure coordinates that have a root mean square deviation equal to or less than the RMSD value specified in Columns 3, 4 or 5 of Table 1 when compared to the structure coordinates of Figure 3, the root mean square deviation being calculated such that the portion of amino acid residues specified in Column 2 of Table 1 of each set of structure coordinates are superimposed and the root mean square deviation is based only on those amino acid residues in the structure coordinates that are also present in the portion of the protein specified in specified in Column 1 of Table 1;

a working memory for storing instructions for processing the machine-readable data;

a central-processing unit coupled to the working memory and to the machine-readable data storage medium, for processing the machine-readable data into the three-dimensional representation; and

an output hardware coupled to the central processing unit, for receiving the three dimensional representation.

[00138] Another embodiment of this invention provides a machine-readable data storage medium, comprising a data storage material encoded with machine readable data which, when

used by a machine programmed with instructions for using said data, displays a graphical three-dimensional representation comprising CatS or a portion or variant thereof.

[00139] In various variations, the machine readable data comprises data for representing a protein based on structure coordinates where the structure coordinates have a root mean square deviation equal to or less than the RMSD value specified in Columns 3, 4 or 5 of Table 1 when compared to the structure coordinates of Figure 3, the root mean square deviation being calculated such that the portion of amino acid residues specified in Column 2 of Table 1 of each set of structure coordinates are superimposed and the root mean square deviation is based only on those amino acid residues in the structure coordinates that are also present in the portion of the protein specified in specified in Column 1 of Table 1.

[00140] According to another embodiment, the machine-readable data storage medium comprises a data storage material encoded with a first set of machine readable data which comprises the Fourier transform of structure coordinates that have a root mean square deviation equal to or less than the RMSD value specified in Columns 3, 4 or 5 of Table 1 when compared to the structure coordinates of Figure 3, the root mean square deviation being calculated such that the portion of amino acid residues specified in Column 2 of Table 1 of each set of structure coordinates are superimposed and the root mean square deviation is based only on those amino acid residues in the structure coordinates that are also present in the portion of the protein specified in specified in Column 1 of Table 1, and which, when using a machine programmed with instructions for using said data, can be combined with a second set of machine readable data comprising the X-ray diffraction pattern of another molecule or molecular complex to determine at least a portion of the structure coordinates corresponding to the second set of machine readable data. For example, the Fourier transform of the structure coordinates set forth in Figure 3 may be used to determine at least a portion of the structure coordinates of other CatS-like enzymes, and isoforms of CatS.

[00141] Optionally, a computer system is provided in combination with the machine-readable data storage medium provided herein. In one embodiment, the computer system comprises a working memory for storing instructions for processing the machine-readable data; a processing unit coupled to the working memory and to the machine-readable data storage medium, for processing the machine-readable data into the three-dimensional representation; and

an output hardware coupled to the processing unit, for receiving the three-dimensional representation.

[00142] Figure 6 illustrates an example of a computer system that may be used in combination with storage media according to the present invention. As illustrated, the computer system 10 includes a computer 11 comprising a central processing unit ("CPU") 20, a working memory 22 which may be, e.g., RAM (random-access memory) or "core" memory, mass storage memory 24 (such as one or more disk drives or CD-ROM drives), one or more cathode-ray tube ("CRT") display terminals 26, one or more keyboards 28, one or more input lines 30, and one or more output lines 40, all of which are interconnected by a conventional bi-directional system bus 50.

[00143] Input hardware 36, coupled to computer 11 by input lines 30, may be implemented in a variety of ways. For example, machine-readable data of this invention may be inputted via the use of a modem or modems 32 connected by a telephone line or dedicated data line 34. Alternatively or additionally, the input hardware 36 may comprise CD-ROM drives or disk drives 24. In conjunction with display terminal 26, keyboard 28 may also be used as an input device.

[00144] Conventional devices may, similarly implement output hardware 46, coupled to computer 11 by output lines 40. By way of example, output hardware 46 may include CRT display terminal 26 for displaying a graphical representation of a binding pocket of this invention using a program such as MOE as described herein. Output hardware might also include a printer 42, so that hard copy output may be produced, or a disk drive 24, to store system output for later use.

[00145] In operation, CPU 20 coordinates the use of the various input and output devices 36, 46 coordinates data accesses from mass storage 24 and accesses to and from working memory 22, and determines the sequence of data processing steps. A number of programs may be used to process the machine-readable data of this invention. Such programs are discussed in reference to using the three dimensional structure of CatS described herein.

[00146] The storage medium encoded with machine-readable data according to the present invention can be any conventional data storage device known in the art. For example, the storage medium can be a conventional floppy diskette or hard disk. The storage medium can also be an optically readable data storage medium, such as a CD-ROM or a DVD-ROM, or a

rewritable medium such as a magneto-optical disk that is optically readable and magneto-optically writable.

8. Uses of the Three Dimensional Structure of CatS

[00147] The three-dimensional crystal structure of the present invention may be used to identify CatS binding sites, be used as a molecular replacement model to solve the structure of unknown crystallized proteins, to design mutants having desirable binding properties, and ultimately, to design, characterize, identify entities capable of interacting with CatS and other structurally similar proteins as well as other uses that would be recognized by one of ordinary skill in the art. Such entities may be chemical entities or proteins. The term "chemical entity", as used herein, refers to chemical compounds, complexes of at least two chemical compounds, and fragments of such compounds.

[00148] The CatS structure coordinates provided herein are useful for screening and identifying drugs that inhibit CatS and other structurally similar proteins. For example, the structure encoded by the data may be computationally evaluated for its ability to associate with putative substrates or ligands. Such compounds that associate with CatS may inhibit CatS, and are potential drug candidates. Additionally or alternatively, the structure encoded by the data may be displayed in a graphical three-dimensional representation on a computer screen. This allows visual inspection of the structure, as well as visual inspection of the structure's association with the compounds.

[00149] Thus, according to another embodiment of the present invention, a method is provided for evaluating the potential of an entity to associate with CatS or a fragment or variant thereof by using all or a portion of the structure coordinates provided in Figure 3 or functional equivalents thereof. A method is also provided for evaluating the potential of an entity to associate with CatS or a fragment or variant thereof by using structure coordinates similar to all or a portion of the structure coordinates provided in Figure 3 or functional equivalents thereof.

[00150] The method may optionally comprise the steps of: creating a computer model of all or a portion of a protein structure (e.g., a binding pocket) using structure coordinates according to the present invention; performing a fitting operation between the entity and the computer model; and analyzing the results of the fitting operation to quantify the association between the entity and the model. The portion of the protein structure used optionally comprises

all of the amino acids listed in Tables 2, 3 and 4 that are present in the structure coordinates being used.

[00151] It is noted that the computer model may not necessarily directly use the structure coordinates. Rather, a computer model can be formed that defines a surface contour that is the same or similar to the surface contour defined by the structure coordinates.

[00152] The structure coordinates provided herein can also be utilized in a method for identifying a ligand (e.g., entities capable of associating with a protein) of a protein comprising an CatS-like binding pocket. One embodiment of the method comprises: using all or a portion of the structure coordinates provided herein to generate a three-dimensional structure of an CatS-like binding pocket; employing the three-dimensional structure to design or select a potential ligand; synthesizing the potential ligand; and contacting the synthesized potential ligand with a protein comprising an CatS-like binding pocket to determine the ability of the potential ligand to interact with protein. According to this method, the structure coordinates used may have a root mean square deviation equal to or less than the RMSD values specified in Columns 3, 4 or 5 of Table 1 when compared to the structure coordinates of Figure 3 according to the RMSD calculation method set forth herein, provided that the portion of amino acid residues specified in Column 2 of Table 1 of each set of structure coordinates are superimposed and the root mean square deviation is calculated based only on those amino acid residues in the structure coordinates that are also present in the portion of the protein specified in Column 1 of Table 1. The portion of the protein structure used optionally comprises all of the amino acids listed in Tables 2, 3, and/or 4 that are present.

[00153] As noted previously, the three-dimensional structure of an CatS-like binding pocket need not be generated directly from structure coordinates. Rather, a computer model can be formed that defines a surface contour that is the same or similar to the surface contour defined by the structure coordinates.

[00154] A method is also provided for evaluating the ability of an entity, such as a compound or a protein to associate with an CatS-like binding pocket, the method comprising: constructing a computer model of a binding pocket defined by structure coordinates that have a root mean square deviation equal to or less than the RMSD value specified in Columns 3, 4 or 5 of Table 1 when compared to the structure coordinates of Figure 3, the root mean square deviation being calculated such that the portion of amino acid residues specified in Column 2 of

Table 1 of each set of structure coordinates are superimposed and the root mean square deviation is based only on those amino acid residues in the structure coordinates that are also present in the portion of the protein specified in specified in Column 1 of Table 1; selecting an entity to be evaluated by a method selected from the group consisting of (i) assembling molecular fragments into the entity, (ii) selecting an entity from a small molecule database, (iii) *de novo* ligand design of the entity, and (iv) modifying a known ligand for CatS, or a portion thereof; performing a fitting program operation between computer models of the entity to be evaluated and the binding pocket in order to provide an energy-minimized configuration of the entity in the binding pocket; and evaluating the results of the fitting operation to quantify the association between the entity and the binding pocket model in order to evaluate the ability of the entity to associate with the said binding pocket.

[00155] The computer model of a binding pocket used in this embodiment need not be generated directly from structure coordinates. Rather, a computer model can be formed that defines a surface contour that is the same or similar to the surface contour defined by the structure coordinates.

[00156] Also according to the method, the method may further include synthesizing the entity; and contacting a protein having an CatS-like binding pocket with the synthesized entity.

[00157] With the structure provided herein, the present invention for the first time permits the use of molecular design techniques to identify, select or design potential inhibitors of CatS, based on the structure of an CatS-like binding pocket. Such a predictive model is valuable in light of the high costs associated with the preparation and testing of the many diverse compounds that may possibly bind to the CatS protein.

[00158] According to this invention, a potential CatS inhibitor may now be evaluated for its ability to bind an CatS-like binding pocket prior to its actual synthesis and testing. If a proposed entity is predicted to have insufficient interaction or association with the binding pocket, preparation and testing of the entity can be obviated. However, if the computer modeling indicates a strong interaction, the entity may then be obtained and tested for its ability to bind.

[00159] A potential inhibitor of an CatS-like binding pocket may be computationally evaluated using a series of steps in which chemical entities or fragments are screened and selected for their ability to associate with the CatS-like binding pockets.

[00160] One skilled in the art may use one of several methods to screen entities (whether chemical or protein) for their ability to associate with an CatS-like binding pocket. This process may begin by visual inspection of, for example, an CatS-like binding pocket on a computer screen based on the CatS structure coordinates in Figure 3 or other coordinates which define a similar shape generated from the machine-readable storage medium. Selected fragments or chemical entities may then be positioned in a variety of orientations, or docked, within that binding pocket as defined above. Docking may be accomplished using software such as Quanta and Sybyl, followed by energy minimization and molecular dynamics with standard molecular mechanics force fields, such as CHARMM and AMBER.

[00161] Specialized computer programs may also assist in the process of selecting entities. These include: GRID (P. J. Goodford, "A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules", J. Med. Chem., 28, pp. 849-857 (1985)). GRID is available from Oxford University, Oxford, UK; MCSS (A. Miranker et al., "Functionality Maps of Binding Sites: A Multiple Copy Simultaneous Search Method." Proteins: Structure, Function and Genetics, 11, pp. 29-34 (1991)). MCSS is available from Molecular Simulations, San Diego, Calif.; AUTODOCK (D. S. Goodsell et al., "Automated Docking of Substrates to Proteins by Simulated Annealing", Proteins: Structure, Function, and Genetics, 8, pp. 195-202 (1990)). AUTODOCK is available from Scripps Research Institute, La Jolla, Calif.; & DOCK (I. D. Kuntz et al., "A Geometric Approach to Macromolecule-Ligand Interactions", J. Mol. Biol., 161, pp. 269-288 (1982)). DOCK is available from University of California, San Francisco, Calif.

[00162] Once suitable entities have been selected, they can be designed or assembled. Assembly may be preceded by visual inspection of the relationship of the fragments to each other on the three-dimensional image displayed on a computer screen in relation to the structure coordinates of CatS. This may then be followed by manual model building using software such as MOE, QUANTA or Sybyl [Tripos Associates, St. Louis, Mo].

[00163] Useful programs to aid one of skill in the art in connecting the individual chemical entities or fragments include: CAVEAT (P. A. Bartlett et al, "CAVEAT: A Program to Facilitate the Structure-Derived Design of Biologically Active Molecules", in "Molecular Recognition in Chemical and Biological Problems", Special Pub., Royal Chem. Soc., 78, pp. 182-196 (1989); G. Lauri and P. A. Bartlett, "CAVEAT: a Program to Facilitate the Design of

Organic Molecules", J. Comput. Aided Mol. Des., 8, pp. 51-66 (1994)). CAVEAT is available from the University of California, Berkeley, Calif.; 3D Database systems such as ISIS (MDL Information Systems, San Leandro, Calif.). This area is reviewed in Y. C. Martin, "3D Database Searching in Drug Design", J. Med. Chem., 35, pp. 2145-2154 (1992); HOOK (M. B. Eisen et al, "HOOK: A Program for Finding Novel Molecular Architectures that Satisfy the Chemical and Steric Requirements of a Macromolecule Binding Site", Proteins: Struct., Funct., Genet., 19, pp. 199-221 (1994). HOOK is available from Molecular Simulations, San Diego, Calif.

[00164] Instead of proceeding to build an inhibitor of an CatS-like binding pocket in a step-wise fashion one fragment or entity at a time as described above, inhibitory or other CatS binding compounds may be designed as a whole or "*de novo*" using either an empty binding site or optionally including some portion(s) of a known inhibitor(s). There are many *de novo* ligand design methods including: LUDI (H.-J. Bohm, "The Computer Program LUDI: A New Method for the De Novo Design of Enzyme Inhibitors", J. Comp. Aid. Molec. Design, 6, pp. 61-78 (1992)). LUDI is available from Molecular Simulations Incorporated, San Diego, Calif.; LEGEND (Y. Nishibata et al., Tetrahedron, 47, p. 8985 (1991)). LEGEND is available from Molecular Simulations Incorporated, San Diego, Calif.; LEAPFROG (available from Tripos Associates, St. Louis, Mo.); & SPROUT (V. Gillet et al, "SPROUT: A Program for Structure Generation)", J. Comput. Aided Mol. Design, 7, pp. 127-153 (1993)). SPROUT is available from the University of Leeds, UK.

[00165] Other molecular modeling techniques may also be employed in accordance with this invention (see, e.g., Cohen et al., "Molecular Modeling Software and Methods for Medicinal Chemistry, J. Med. Chem., 33, pp. 883-894 (1990); see also, M. A. Navia and M. A. Murcko, "The Use of Structural Information in Drug Design", Current Opinions in Structural Biology, 2, pp. 202-210 (1992); L. M. Balbes et al., "A Perspective of Modern Methods in Computer-Aided Drug Design", in Reviews in Computational Chemistry, Vol. 5, K. B. Lipkowitz and D. B. Boyd, Eds., VCH, New York, pp. 337-380 (1994); see also, W. C. Guida, "Software For Structure-Based Drug Design", Curr. Opin. Struct. Biology, 4, pp. 777-781 (1994)).

[00166] Once an entity has been designed or selected, for example, by the above methods, the efficiency with which that entity may bind to an CatS binding pocket may be tested and optimized by computational evaluation. For example, an effective CatS binding pocket inhibitor preferably demonstrates a relatively small difference in energy between its bound and free states

(i.e., a small deformation energy of binding). Thus, the most efficient CatS binding pocket inhibitors should preferably be designed with deformation energy of binding of not greater than about 10 kcal/mole, more preferably, not greater than 7 kcal/mole. CatS binding pocket inhibitors may interact with the binding pocket in more than one of multiple conformations that are similar in overall binding energy. In those cases, the deformation energy of binding is taken to be the difference between the energy of the free entity and the average energy of the conformations observed when the inhibitor binds to the protein.

[00167] An entity designed or selected as binding to an CatS binding pocket may be further computationally optimized so that in its bound state it would preferably lack repulsive electrostatic interaction with the target enzyme and with the surrounding water molecules. Such non-complementary electrostatic interactions include repulsive charge-charge, dipole-dipole and charge-dipole interactions.

[00168] Specific computer software is available in the art to evaluate compound deformation energy and electrostatic interactions. Examples of programs designed for such uses include: Gaussian 94, revision C (M. J. Frisch, Gaussian, Inc., Pittsburgh, Pa. COPYRGT.1995); AMBER, version 4.1 (P. A. Kollman, University of California at San Francisco, COPYRGT 1995); QUANTA/CHARMM (Molecular Simulations, Inc., San Diego, Calif. COPYRGT.1995); Insight II/Discover (Molecular Simulations, Inc., San Diego, Calif. COPYRGT.1995); DelPhi (Molecular Simulations, Inc., San Diego, Calif. COPYRGT.1995); and AMSOL (Quantum Chemistry Program Exchange, Indiana University). These programs may be implemented, for instance, using a Silicon Graphics workstation such as an Indigo.sup.2 with "IMPACT" graphics. Other hardware systems and software packages will be known to those skilled in the art.

[00169] Another approach provided by this invention, is the computational screening of small molecule databases for chemical entities or compounds that can bind in whole, or in part, to an CatS binding pocket. In this screening, the quality of fit of such entities to the binding site may be judged either by shape complementarities or by estimated interaction energy [E. C. Meng et al., J. Comp. Chem., 13, 505-524 (1992)].

[00170] According to another embodiment, the invention provides compounds that associate with an CatS-like binding pocket produced or identified by various methods set forth above.

[00171] The structure coordinates set forth in Figure 3 can also be used to aid in obtaining structural information about another crystallized molecule or molecular complex. This may be achieved by any of a number of well-known techniques, including molecular replacement.

[00172] For example, a method is also provided for utilizing molecular replacement to obtain structural information about a protein whose structure is unknown comprising the steps of: generating an X-ray diffraction pattern of a crystal of the protein whose structure is unknown; generating a three-dimensional electron density map of the protein whose structure is unknown from the X-ray diffraction pattern by using at least a portion of the structure coordinates set forth in Figure 3 as a molecular replacement model.

[00173] By using molecular replacement, all or part of the structure coordinates of the CatS provided by this invention (and set forth in Figure 3) can be used to determine the structure of another crystallized molecule or molecular complex more quickly and efficiently than attempting an *ab initio* structure determination. One particular use includes use with other structurally similar proteins. Molecular replacement provides an accurate estimation of the phases for an unknown structure. Phases are a factor in equations used to solve crystal structures that cannot be determined directly. Obtaining accurate values for the phases, by methods other than molecular replacement, is a time-consuming process that involves iterative cycles of approximations and refinements and greatly hinders the solution of crystal structures. However, when the crystal structure of a protein containing at least a homologous portion has been solved, the phases from the known structure provide a satisfactory estimate of the phases for the unknown structure.

[00174] Thus, this method involves generating a preliminary model of a molecule or molecular complex whose structure coordinates are unknown, by orienting and positioning the relevant portion of CatS according to Figure 3 within the unit cell of the crystal of the unknown molecule or molecular complex so as best to account for the observed X-ray diffraction pattern of the crystal of the molecule or molecular complex whose structure is unknown. Phases can then be calculated from this model and combined with the observed X-ray diffraction pattern amplitudes to generate an electron density map of the structure whose coordinates are unknown. This, in turn, can be subjected to any well-known model building and structure refinement techniques to provide a final, accurate structure of the unknown crystallized molecule or molecular complex [E. Lattman, "Use of the Rotation and Translation Functions", in Meth.

Enzymol., 115, pp. 55-77 (1985); M. G. Rossmann, ed., "The Molecular Replacement Method", Int. Sci. Rev. Ser., No. 13, Gordon & Breach, New York (1972)].

[00175] The structure of any portion of any crystallized molecule or molecular complex that is sufficiently homologous to any portion of CatS can be resolved by this method.

[00176] In one embodiment, the method of molecular replacement is utilized to obtain structural information about the present invention and any other CatS-like molecule. The structure coordinates of CatS, as provided by this invention, are particularly useful in solving the structure of other isoforms of CatS or CatS complexes.

[00177] The structure coordinates of CatS as provided by this invention are useful in solving the structure of CatS variants that have amino acid substitutions, additions and/or deletions (referred to collectively as "CatS mutants", as compared to naturally occurring CatS). These CatS mutants may optionally be crystallized in co-complex with a ligand, such as an inhibitor, substrate analogue or a suicide substrate. The crystal structures of a series of such complexes may then be solved by molecular replacement and compared with that of CatS. Potential sites for modification within the various binding sites of the enzyme may thus be identified. This information provides an additional tool for determining the most efficient binding interactions such as, for example, increased hydrophobic interactions, between CatS and a ligand. It is noted that the ligand may be the protein's natural ligand or may be a potential agonist or antagonist of a protein.

[00178] All of the complexes referred to above may be studied using well-known X-ray diffraction techniques and may be refined versus 1.5-3Å resolution X-ray data to an R value of about 0.22 or less using computer software, such as X-PLOR [Yale University, COPYRIGHT.1992, distributed by Molecular Simulations, Inc.; see, e.g., Blundell & Johnson, *supra*; Meth. Enzymol., Vol. 114 & 115, H. W. Wyckoff et al., eds., Academic Press (1985)]. This information may thus be used to optimize known CatS inhibitors, and more importantly, to design new CatS inhibitors.

[00179] The structure coordinates described above may also be used to derive the dihedral angles, phi and psi, that define the conformation of the amino acids in the protein backbone. As will be understood by those skilled in the art, the phi_n angle refers to the rotation around the bond between the alpha-carbon and the nitrogen, and the psi_n angle refers to the rotation around the bond between the carbonyl carbon and the alpha-carbon. The subscript "n" identifies the amino

acid whose conformation is being described [for a general reference, see Blundell and Johnson, Protein Crystallography, Academic Press, London, 1976].

9. Uses of the Crystal and Diffraction Pattern of CatS

[00180] Crystals, crystallization conditions and the diffraction pattern of CatS that can be generated from the crystals also have a range of uses. One particular use relates to screening entities that are not known ligands of CatS for their ability to bind to CatS. For example, with the availability of crystallization conditions, crystals and diffraction patterns of CatS provided according to the present invention, it is possible to take a crystal of CatS; expose the crystal to one or more entities that may be a ligand of CatS; and determine whether a ligand/ CatS complex is formed. The crystals of CatS may be exposed to potential ligands by various methods, including but not limited to, soaking a crystal in a solution of one or more potential ligands or co-crystallizing CatS in the presence of one or more potential ligands. Given the structure coordinates provided herein, once a ligand complex is formed, the structure coordinates can be used as a model in molecular replacement in order to determine the structure of the ligand complex.

[00181] Once one or more ligands are identified, structural information from the ligand/ CatS complex(es) may be used to design new ligands that bind tighter, bind more specifically, have better biological activity or have better safety profile than known ligands.

[00182] In one embodiment, a method is provided for identifying a ligand that binds to CatS comprising: (a) attempting to crystallize a protein that comprises a sequence wherein at least a portion of the sequence has 55%, 65%, 75%, 85%, 90%, 95%, 97%, 99% or greater identity with SEQ. ID No. 3 in the presence of one or more entities; (b) if crystals of the protein are obtained in step (a), obtaining an X-ray diffraction pattern of the protein crystal; and (c) determining whether a ligand/protein complex was formed by comparing an X-ray diffraction pattern of a crystal of the protein formed in the absence of the one or more entities to the crystal formed in the presence of the one or more entities.

[00183] In another embodiment, a method is provided for identifying a ligand that binds to CatS comprising: soaking a crystal of a protein wherein at least a portion of the protein has 55%, 65%, 75%, 85%, 90%, 95%, 97%, 99% or greater identity with SEQ. ID No. 3 with one or more entities; determining whether a ligand/protein complex was formed by comparing an X-ray

diffraction pattern of a crystal of the protein that has not been soaked with the one or more entities to the crystal that has been soaked with the one or more entities.

[00184] Optionally, the method may further comprise converting the diffraction patterns into electron density maps using phases of the protein crystal and comparing the electron density maps.

[00185] Libraries of "shape-diverse" compounds may optionally be used to allow direct identification of the ligand-receptor complex even when the ligand is exposed as part of a mixture. According to this variation, the need for time-consuming de-convolution of a hit from the mixture is avoided. More specifically, the calculated electron density function reveals the binding event, identifies the bound compound and provides a detailed 3-D structure of the ligand-receptor complex. Once a hit is found, one may optionally also screen a number of analogs or derivatives of the hit for tighter binding or better biological activity by traditional screening methods. The hit and information about the structure of the target may also be used to develop analogs or derivatives with tighter binding or better biological activity. It is noted that the ligand-CatS complex may optionally be exposed to additional iterations of potential ligands so that two or more hits can be linked together to make a more potent ligand. Screening for potential ligands by co-crystallization and/or soaking is further described in U.S. Patent No. 6,297,021, which is incorporated herein by reference.

EXAMPLES

Example 1. Expression and Purification of CatS

[00186] This example describes the expression of CatS. It should be noted that a variety of other expression systems and hosts are also suitable for the expression of CatS, as would be readily appreciated by one of skill in the art.

[00187] The portion of the gene encoding residues 1-331 (from SEQ. ID No. 1) which corresponds to the entire sequence of human CatS was amplified by PCR and cloned into a modified pFastbac vector (Invitrogen) with a 6-histidine tag at the C-terminus. This DNA sequence is presented in Figure 1 as SEQ. ID No. 2.

[00188] Expression in this vector generated a fusion of CatS residues 1-340 with a C-terminal 6x-histidine tag, the amino acid sequence of which is shown in Figure 1 as SEQ. ID. No. 4. Recombinant baculoviruses incorporating the CatS constructs were generated by

transposition using the Bac-to-Bac system (Invitrogen). High-titer viral stocks were generated by infection of *Spodoptera frugiperda* Sf9 cells and the expression of recombinant protein was carried out by infection of *Trichoplusia ni* Hi5 cells (Invitrogen) in 10L Wave Bioreactors (Wave Biotech).

[00189] Most of the protein was secreted into the media. The cell supernatant was concentrated, and diafiltered by cross flow ultrafiltration. The protein in the supernatant was purified by passage over ProBond (InVitrogen) resin. After activation was achieved by altering the pH to low pH, E64 was added and cation exchange chromatography was used to isolate the CatS-E64 complex. The CatS protein purity as determined on denaturing SDS-PAGE gel was 90-95%. CatS was concentrated to a final concentration of 8.5 mg/ml and stored at 4°C in a buffer containing 25 mM Sodium Acetate, pH 5.5, 150 mM NaCl, 1.5 mM benzamidine and a five-fold molar excess of E64.

Example 2. Crystallization of CatS-E64 complex

[00190] This example describes the crystallization of the CatS-E64 complex. It is noted that the precise crystallization conditions used may be further varied, for example by performing a fine screen based on these crystallization conditions.

[00191] CatS protein samples were incubated with 1.5mM benzamidine and a fivefold molar excess of E64 before setting crystallization trials. Crystals were obtained after an extensive and broad screen of conditions, followed by optimization. Diffraction quality crystals were grown as in 100nL sitting droplets using the vapor diffusion method. 50nL comprising the CatS-E64 complex (8.5 mg/ml) was mixed with 50nL from a reservoir solution (100µL) comprising 0.2M citrate/citric acid pH=5.0, 16% (w/v) PEG 4000 and 5% (w/v) PEG 200. The resulting solution was incubated over a period of one week at 4°C.

[00192] Crystals typically appeared after 48 hours and grew to a maximum size within 72 hours. Single crystals were transferred, briefly, into a cryoprotecting solution containing the reservoir solution supplemented with 30% v/v glycerol. Crystals were then flash frozen by immersion in liquid nitrogen and then stored under liquid nitrogen. A crystal of CatS-E64 complex produced as described is illustrated in Figure 2.

[00193] While the present invention is disclosed with reference to certain embodiments and examples detailed above, it is to be understood that these embodiments and examples are

intended to be illustrative rather than limiting, as it is contemplated that modifications will readily occur to those skilled in the art, which modifications are intended to be within the scope of the invention and the appended claims. All patents, papers, and books cited in this application are incorporated herein in their entirety.